Model Selection for the Competing Risks Model With and Without Masking

Radu V. Craiu

Department of Statistics, University of Toronto,

100 St. George Street, Toronto, ON M5S 3G3, Canada

craiu@utstat.toronto.edu

Thomas C. M. Lee

Department of Statistics, Colorado State University,

Fort Collins, CO 80523-1877, U.S.A.

tlee@stat.colostate.edu

January 19, 2005

Abstract

The competing risks model is useful in settings in which individuals (or units) may die (or fail) due to a number of different causes. It can also be the case that for some of the items the failure cause is known only up to a subgroup of all causes in which case we say that the failure is group masked. A widely used approach for competing risks data with and without masking involves the specification of cause-specific hazard rates. Often, due to the availability of likelihood methods for estimation and testing, piecewise constant hazards are used. The piecewise constant rates also offer model flexibility and

computational convenience. However, for such piecewise constant hazard models the choice of the end points for each interval on which the hazards are constant is usually a subjective one. In this paper we discuss and propose the use of model selection methods that are data driven and automatic. We compare three model selection procedures, based on the Minimum Description Length (MDL) principle, the Bayes Information Criterion (BIC) and the Akaike Information Criterion (AIC). A fast splitting algorithm is the computational tool used to select among an enormous number of possible models. We test the effectiveness of the methods via numerical studies, including a real dataset with masked failure causes.

KEY WORDS: Akaike information criterion; Bayesian information criterion; Code length; Competing risks; EM algorithm; Group masked cause; Minimum description length principle; Missing data; Model selection; Piecewise constant hazard

1 Introduction

In many survival data studies it is the case that the individuals under study can experience any one of J types of failure. Consequently, each individual/item under study has associated with it J potential failure times, one for each possible failure. Obviously, in practice, only one of the potential failure times is observed unless the item is right censored, i.e., it does not fail before the end of the study, in which case no failure time is observed. The competing risks problem involves the estimation of failure rates for each type of failure. An additional complication arises when a subset of the individuals have a cause of failure that is only known to belong to a certain subset of all possible causes, in other words, their cause of failure is *group masked*. In practice one possibility is to conduct a second-stage analysis, such as autopsy, in which the true cause can be uniquely determined. In fact, inference is possible even if not all items are subjected to a second stage analysis as we discuss in Section 2.2.

Examples of failure data obtained under a competing risks model are abundant in the literature and range from survival analysis studies in biostatistics to applications of reliability in engineering to risk models in actuarial science. For instance, Gaynor et al. (1993) and Barret et al. (1989) discuss the importance of estimating the probability of death due to cancer relapse after treatment versus the probability of death due to treatment-related complications, providing an example where the competing risks are not acting independently. In reliability studies, Sun and Tiwari (1997) analyze the failure times of small electrical appliances that may fail due to two competing risks while Taylor (1994) uses competing risks to model the probability distribution of the tensile strength of certain materials known

to contain two or more subpopulations of flaw types. Lapidus et al. (1994) present a study of motorcycle fatalities in which 40% of the death certificates have missing information.

Parametric analyses of the competing risks model are proposed by Hoel (1972), Moeschberger and David (1978), Lagakos (1977), and Prentice et al. (1978). Cause-specific hazard functions are used in nonparametric estimation by Nelson (1969), Aalen (1978), and Crowder (2001). Semiparametric methods based on proportional hazards models were discussed by Holt (1978), Kalbfleisch and Prentice (2002, Ch. 8) and Lawless (2003, Ch. 9).

For the competing risks model in which a subset of all items have masked causes of failure, some authors have derived semi- and nonparametric inference procedures in the case with two failure causes and no second-stage analysis which often occurs in carcinogenicity bioassays: Dinse (1986) proposed nonparametric maximum likelihood estimators of prevalence and mortality; Goetghebeur and Ryan (1990) derived a modified log-rank test for comparing the survival of populations, which they later extended to proportional hazards regression (Goetghebeur and Ryan, 1995); Racine-Poon and Hoel (1984) considered inference for this model when a probability of death from each missing cause is provided by a pathologist; and Kodell and Chen (1987) tackled the problem via the EM algorithm. In the case of a general number of failure causes and availability of second-stage analysis data, Flehinger, Reiser and Yashchin (1998, 2002) propose maximum likelihood estimation under a model with nonparametric proportional cause-specific hazards (Flehinger et al., 1998) and a model with completely parametric cause-specific hazards (Flehinger et al., 2002). Craiu and Duchesne (2004a) propose a semiparametric model with piecewise constant cause-specific hazard functions which presents robust properties and can be used in most situations in which some second stage data is available. The same approach can be used in order to

integrate prior knowledge about the failure process using a Bayesian analysis in which the data augmentation algorithm is used for computation (Craiu and Duchesne, 2004b).

The model with piecewise constant cause-specific hazard functions achieves a good balance between flexibility and accuracy on one side and computational feasibility on the other side. In addition, while essentially non-parametric, these models allow for likelihood based methods for estimation and testing to be used (Craiu and Duchesne, 2004a; Lawless, 2003; He and Lawless, 2003). Generally the end-points of the intervals that define the piecewise constant hazard functions (henceforth simply called "the intervals") are chosen by each researcher based on their past experience or intuition regarding the failure process.

The main contribution of this paper is the exploration of more objective and data-specific criteria for automatically choosing the intervals. To the best of our knowledge, this is the first time that such a study is conducted. In particular, we focus on three widely used model selection criteria, the Minimum Description Length (MDL) principle, the Bayesian Information Criterion (BIC) and a small sample version of the Akaike Information Criterion (AICC). None of the three criteria is uniformly optimal. The AICC and MDL seem to perform better if the number of intervals (relative to the number of observations) is large, while the BIC is slightly better than the MDL in situations in which only a small number of intervals is needed, with AICC lagging behind in this case. However, for situations in which the statistician does not have any knowledge regarding the number of intervals we recommend the MDL criterion since, on average, it performs the best.

In the next section we describe the data and the likelihood methods used for estimation. We also provide some theoretical justification of the importance of correct selection of the end points for each interval. In Section 3 we discuss in detail the three criteria used for

model selection and show how they can be applied to the competing risks problem. Real examples and a simulation study are performed in Section 4 to illustrate the efficiency of the model selection procedure. We close with conclusions and ideas for further improvements.

2 Data and Models

2.1 Competing risks without masking

In the competing risks model with unmasked data, assume that there are J possible failure causes and N items are observed between time $t_0 = 0$ and $t_{\max}$, the time when the study is stopped. To each item i which has failed at a time $t_i \in [t_0, t_{\max}]$ it corresponds a pair (c_i, t_i) in which c_i is equal to the cause of failure, i.e. $c_i = j$ if the jth cause is responsible for the failure. In general we refer to (c_i, t_i) as one realization of a bivariate random variable (C, T). For those items which have not failed during the observation period $[t_0, t_{\max}]$, that is for those items that are right censored, neither c_i nor t_i is observed. In our model selection procedure we will use only the uncensored items since the censored items do not contain any information regarding the cut points of the intervals. Practically, this implies that we ignore in the likelihood the terms involving censored items. It should be noted that such assumption should not be carried over in the estimation phase of the study, once the model is selected.

The dependence between T and C is usually specified using the cause-specific hazard rates

$$\lambda_j(t) = \lim_{h \downarrow 0} \frac{\Pr(t < T \leq t + h, C = j | T \geq t)}{h}, \quad j = 1, \ldots, J. \tag{1}$$

From equation (1) it follows that the marginal hazard function for T is $\lambda(t) = \sum_{j=1}^{J} \lambda_j(t)$

and the marginal survivor function for T is $S(t) = \Pr(T > t) = \exp\left\{-\int_0^t \sum_{j=1}^J \lambda_j(u)\, du\right\}$. The cumulative incidence functions are $F_j(t) = \Pr(T \leq t, C = j) = \int_0^t \lambda_j(t) S(t)\, dt$.

We define each cause-specific hazard rate to be a piecewise constant function, that is, we partition the interval $[0, t_{\max}]$ into K disjoint intervals $(a_{k-1}, a_k]$ so that

$$\lambda_j(t) = \sum_{k=1}^{K} \lambda_{jk} 1_k(t), \tag{2}$$

where $0 = a_0 < a_1 < \cdots < a_K = t_{\max}$, and $1_k(t)$ is the indicator that $t \in (a_{k-1}, a_k]$. Note that in (2) we have made the implicit assumption that the cut points $a_0, \ldots, a_K$ are the same for all failure causes. This assumption is not necessary in order to carry out the model selection procedure presented here. However, the use of common cut points simplifies the notation and the understanding of the ideas. In addition, the model with equal intervals across causes encompasses the proportional hazards model with piecewise hazards in which $\lambda_j(t) = \sum_{k=1}^K r_{jk} \lambda(t) 1_k(t)$ and $\sum_j r_{jk} = 1$. For the remaining of the paper we will assume the model defined by (2).

The likelihood function is then proportional to

$$L(\theta) \;=\; \prod_{i=1}^{N} \prod_{j=1}^{J} \left[\left\{ \sum_{k=1}^{K} \lambda_{jk} 1_k(t_i) \right\}^{\delta_{ij}} \exp\left\{ -\int_0^{t_i} \sum_{k=1}^{K} \lambda_{jk} 1_k(u)\, du \right\} \right], \tag{3}$$

where θ is the $(J \times K)$-dimensional vector of parameters $(\lambda_{11}, \ldots, \lambda_{JK})$. It is convenient to introduce, for each item i and for each cause j, the indicator δ_{ij} which is equal to one if item i has failed due to cause j and it is equal to zero otherwise. The maximum likelihood estimate for θ can be obtained from (3) as

$$\hat{\lambda}_{jk} = \frac{\sum_{i=1}^{N} \delta_{ij} 1_k(t_i)}{e_k}, \tag{4}$$

where e_k is the exposure in the interval $(a_{k-1}, a_k]$, i.e. the sum of all the time lived by each

item in this interval. For instance, if item 1 has failure time $t_1 \in (a_{k-1}, a_k]$ and item 2 has failure time $t_2 > a_k$, their contributions to e_k is $t_1 - a_{k-1}$ and $a_k - a_{k-1}$, respectively.

It is clear from (4) that the choice of the cut points is crucial for the estimation of each parameter λ_{jk}. To better understand the impact of misspecification of a_k's on the estimates, consider the following simple example in which the data is generated under a model, M_0 with two competing risks each having constant cause-specific hazards, λ_1 and λ_2 respectively, on the interval $[0, t_{max}]$. However, suppose that we fail to choose this model and instead we work with a model M_1 in which the cut points are $0 = a_0 < a_1 < a_2 = t_{max}$ so that for each cause $j = 1, 2$ the cause-specific hazard is

$$\lambda_j(t) = \lambda_{j1} 1_{(0,a_1]}(t) + \lambda_{j2} 1_{(a_1,t_{max}]}(t). \tag{5}$$

Denote n_{jk} the number of items that died of cause j in the interval $(a_{k-1}, a_k]$ for each $k = 1, 2$.

Under the true model, M_0, $\hat{\lambda}_1 = \frac{n_{11}+n_{12}}{e_1+e_2}$ and under model M_1, $\hat{\lambda}_{11} = \frac{n_{11}}{e_1}$. We prove in the appendix the following result.

Lemma 2.1 *As $N \to \infty$ the following hold:*

i) $\hat{\lambda}_{11}$ and $\hat{\lambda}_1$ converge almost surely to λ_1.

ii) The variance of $\hat{\lambda}_{11}$ is larger than the variance of $\hat{\lambda}_1$.

This holds even for moderate values of N. For example, with a sample size $N = 50$ simulations show that the variance of $\hat{\lambda}_1$ is 25% smaller than the variance of $\hat{\lambda}_{11}$ when $a_1 = t_{max}/2$. The situation discussed above describes a type of error which results only in variance inflation. However, if the original 'true model' has piecewise cause-specific hazards with more than one interval, it is likely that one of the true end-points will be included

inside one of the assumed (misspecified) intervals. In such a case, calculations similar to those above show that the estimates are asymptotically biased and less efficient than the estimates obtained under the true model. We emphasize that the asymptotic results are obtained under the assumption that the size of the sample increases but the true model as well as the specified model intervals remain constant. We must note that in practice it is usually the case that the piecewise constant hazards are just an approximation to the true ones. However, Lemma 2.1 shows that it is possible to increase the efficiency of the estimators for the flat segments of the true hazards if the intervals end points are properly selected. Simulations in Section 4.1 indeed reflect the result of the lemma.

2.2 Competing risks with masking

The simple competing risks model presented before becomes rapidly more complicated once some of the items have unknown failure causes. In particular, we consider here the case when one can narrow down the cause of failure to a group of possible causes, in other words, the item's failure is *group masked*. In addition, we assume that some of the items with a masked failure cause are sent to a second stage analysis for the purpose of determining the exact reason for failure.

Therefore, in the case of masked data, for each item i, there are three possible occurrences: i fails because of cause j_i at time t_i; i fails because of a cause that is not known precisely, but is known to belong to a group of failure causes $g_i \subset \{1, \ldots, J\}$; or i had still not failed by time t_i. Therefore, some of the items will have a masking group instead of a failure cause, and all the items have a failure time. If G is the number of proper groups, i.e. groups that contain more than one element, then the observation for item i is

$(t_i, \gamma_{ig_1}, \ldots, \gamma_{ig_{G+J}}, \delta_{i1}, \ldots, \delta_{iJ})$, where γ_{ig} is the indicator that item i's failure cause was masked to group g at the first stage; if the failure cause is known to be j at the first stage, then we say that it is masked to $g = \{j\}$. Also, δ_{ij} is the indicator that item i's actual failure cause is j. Obviously, if the item is masked in the initial stage and it is not sent for further analysis, the indicators δ_{ij} are not known and we denote by $\mathcal{M}$ the set of all such items.

As a result of masking, in addition to the parameters λ_{jk}, one must consider the *masking probabilities*

$$P_{g|j} = \Pr(\text{cause masked to group } g \text{ at stage } 1 | C = j), \quad j \in g. \tag{6}$$

Of eventual interest to practitioners are the diagnostic probabilities (Flehinger et al., 1998, 2002)

$$\pi_{j|g}(t) = \Pr(\text{actually failed of cause } j | \text{failed at time } t \text{ and failure cause masked in } g).$$

Using Bayes' rule one obtains

$$\pi_{j|g}(t) = \frac{\lambda_j(t) P_{g|j}}{\sum_{l \in g} \lambda_l(t) P_{g|l}}. \tag{7}$$

For such data, Craiu and Duchesne (2004a) have developed an EM algorithm (Dempster et al., 1977) in which the δ_{ij} (for those masked items) are treated as missing data and which allows the estimation of all the parameters of the model. In the model selection procedures used in the following sections, an essential ingredient is the observed likelihood function as well as the estimators for each parameter of interest. We briefly review here the algorithm used for estimation while we refer to Craiu and Duchesne (2004a) for details concerning additional properties such as convergence and variance estimation.

Using equations (1)-(6) we obtain the loglikelihood function under the complete data as follows

$$l_C(\theta) = \sum_{i=1}^{N}\sum_{j=1}^{J}\left\{\left[\delta_{ij}\ln\sum_{k=1}^{K}\lambda_{jk}1_k(t_i) - \sum_{k=1}^{K}\lambda_{jk}\int_0^{t_i}1_k(u)\,du\right] + \delta_{ij}\left[(1-\sum_{g\in\mathcal{G}_j^*}\gamma_{ig})\ln(1-\sum_{g\in\mathcal{G}_j^*}P_{g|j}) + \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln P_{g|j}\right]\right\}, \quad (8)$$

where in this case θ is the vector of parameters λ_{jk} and $P_{g|j}$ for all $1 \le j \le J$, $1 \le k \le K$ and all masking groups g, and $\mathcal{G}_j^*$ is the number of proper masking groups that contain cause j, for all $1 \le j \le J$.

For right-censored observations, the term on the second line of equation (8) vanishes, and hence the γ_{ig} are not needed for right-censored observations. We emphasize again that for the stated purpose of the paper, i.e. the choice of the intervals' limits, we consider that there are no right-censored observations. The EM algorithm consists in the following steps:

Initial step Set $\hat{\lambda}_{jk}^{(0)} = \sum_{i=1}^{N} 1[\delta_{ij} \text{ observed and equal to } 1]/e_k$, $j = 1,\ldots,J$, $k = 1,\ldots,K$ and $\hat{P}_{g|j}^{(0)} = 1/\#\mathcal{G}_j$, $j = 1,\ldots,J$, $g = g_1,\ldots,g_{G+J}$, where $\#$ denotes cardinality and $\mathcal{G}_j$ is the set of all masking groups that contain cause j.

E-step Using (7) compute $E_{\hat{\theta}^{(l-1)}}[\delta_{ij}|OBS]$ as

$$E_\theta[\delta_{ij}|OBS] = \begin{cases} 1, & \text{cause of failure of } i \text{ known to be } j. \\ 0, & \text{cause of failure of } i \text{ known not to be } j. \\ \hat{\pi}_{j|g_i}(t_i), & \text{cause of } i \text{ masked in } g_i \text{ and no stage 2 data for } i. \end{cases} \quad (9)$$

The $\hat{\pi}_{j|g_i}(t_i)$ is computed using (7).

M-step Set

$$\hat{\lambda}_{jk}^{(l)} = \frac{\sum_{i=1}^{N} E_{\hat{\theta}^{(l-1)}}[\delta_{ij}|OBS]\ 1_k(t_i)}{e_k} \quad \text{and} \quad \hat{P}_{g|j}^{(l)} = \frac{\sum_{i=1}^{N} E_{\hat{\theta}^{(l-1)}}[\delta_{ij}|OBS]\ \gamma_{ig}}{\sum_{i=1}^{N} E_{\hat{\theta}^{(l-1)}}[\delta_{ij}|OBS]}. \quad (10)$$

One can notice in (10) that the expression for $\hat{\lambda}_{jk}^{(l)}$ is the same as the one in (4) with the exception of those δ_{ij} which are unknown and who must be replaced by their estimates computed in the E-step.

We would like to note, that the observed log-likelihood can also be computed in this situation since

$$\begin{aligned} l_{OBS}(\theta) &= E_{\theta'}[l_C(\theta)|OBS] - E_{\theta'}[l_{\mathcal{M}}(\theta)|OBS] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{J}\left\{\left[E_{\theta'}[\delta_{ij}|OBS]\ln\sum_{k=1}^{K}\lambda_{jk}1_k(t_i) - \sum_{k=1}^{K}\lambda_{jk}\int_0^{t_i}1_k(u)\,du\right]\right. \\ &\quad \left. + E_{\theta'}[\delta_{ij}|OBS]\left[(1-\sum_{g\in\mathcal{G}_j^*}\gamma_{ig})\ln(1-\sum_{g\in\mathcal{G}_j^*}P_{g|j}) + \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln P_{g|j}\right]\right\} \\ &\quad - \sum_{i\in\mathcal{M}}\sum_{j\in g_i}E_{\theta'}[\delta_{ij}|OBS]\ln\pi_{j|g_i}(t_i). \end{aligned} \quad (11)$$

3 Model Selection Methods

In this section we consider the problem of selecting a "best" fitting model; that is, the problem of choosing a "best" number of intervals K and a "best" combination of the interval end points a_k. Recall that, in our convention, $a_0 = 0$ and $a_K = t_{\max}$, and hence there are $K-1$ end points to be determined. For simplicity, we write $A_K = (a_1, \ldots, a_{K-1})$. Notice that once A_K is specified, unique maximum likelihood estimates for $\theta = (\lambda_{11}, \ldots, \lambda_{JK})$ can be obtained using (4) or via the EM algorithm described in Section 2.2.

We suggest using the following strategy to solve the problem of finding a "best" A_K. First, a model selection principle is applied to **define** a "best" fitting model. Then a fast

splitting algorithm is adopted to practically obtain such a defined "best" fitting model. In the two subsections below, we discuss the use of three different model selection principles for defining a "best" fitting model.

3.1 Akaike and Bayesian Information Criteria

With the Akaike Information Criterion (AIC) the best fitting model is defined as the minimizer of an estimator of the Kullback–Leibler (KL) distance measure between a fitted model and the "true" model (e.g., see Burnham and Anderson, 2002). If r is the number of parameters that need to be estimated in a fitted model, then under some mild regularity conditions it can be shown that such a KL distance estimator is $-2\times$ "maximized log likelihood" $+2r$. Here for a candidate model with K intervals, there are JK λ_{jk}'s to be estimated, and $M = \sum_{h=1}^{G} \#g_h - J$ masking probabilities ($\#g_h$ denotes the cardinality of the masking group g_h). The number of independent parameters is thus $r = JK + M$ and the AIC best fitting model is defined as the one that minimizes

$$\mathrm{AIC}(A_K) = -2l_{OBS}(\theta) + 2(JK + M). \tag{12}$$

It has been known that this criterion is biased when the sample size is small, and, for many problems biased-corrected versions of it have been proposed (e.g., Burnham and Anderson, 2002 and McQuarrie and Tsai, 1998). Such small sample version criteria are often termed AICC, and for the present problem it is given by

$$\mathrm{AICC}(A_K) = -2l_{OBS}(\theta) + 2(JK + M) + \frac{2(JK + M)(JK + M + 1)}{N - JK - M - 1}. \tag{13}$$

Our simulation study suggests that AICC is uniformly better than AIC, and hence AIC is not included in our summary of findings.

The form of the Bayesian Information Criterion (BIC; Schwarz, 1978) is very similar to AIC. Instead of a constant value 2, it replaces the penalty for each parameter with $\log N$. Thus the BIC best fitting model is defined as the one that minimizes

$$\mathrm{BIC}(A_K) = -2l_{OBS}(\theta) + (M + JK)\log N. \tag{14}$$

As stated in Hastie, Tibshirani and Friedman (2002), choosing the model with the minimum BIC value is approximately equivalent to choosing the model with the largest posterior probability with respect to an uniform prior. Since the penalty term for BIC is larger than the one for AIC, it is expected that BIC tends to produce more parsimonious best fitting models than AIC. However, this comparison is less clear with respect to the AICC, especially if the sample size N is not very large. An astute reader will have noticed that, in the case of masked data, the number of masking probabilities, M, can be omitted in the equations (12) and (14) since this number remains the same no matter how many intervals we use. However, this is not the case for AICC, as can be seen from (13), so the number M must be taken into consideration in that case.

3.2 Minimum Description Length Principle

The MDL principle uses ideas from the information theory and signal processing literature and was adapted by Rissanen (1989) as a model selection tool for statisticians. It *defines* the best fitting model as the one that produces the shortest code length of the data. Loosely speaking, the code length of an object can be treated as the amount of memory space that is required to store the object. For details, see Rissanen (1989). See also for examples Hansen and Yu (2001) and Lee (2001) for introductory tutorials to the MDL principle.

One common approach to apply the MDL principle is to split the code length for a set of data into two components: (i) a fitted model plus (ii) the data "conditioned on" the fitted model; i.e., the part in the data that is not explained by the fitted model. For the present problem, a fitted model can be specified by A_K and the maximum likelihood estimate $\hat{\theta} = (\hat{\lambda}_{11}, \ldots, \hat{\lambda}_{JK})$ for θ. We choose to omit the number of masking parameters, M, from the criterion since this number remains the same across models with different intervals. If $CL(z)$ denotes the code length of the object z, we have the following decomposition:

$$\begin{aligned} CL(\text{“data”}) &= CL(A_K, \hat{\theta}) + CL(\text{“data”}|A_K, \hat{\theta}) \\ &= CL(A_K) + CL(\hat{\theta}) + CL(\text{“data”}|A_K, \hat{\theta}). \end{aligned}$$

Now the task is to find an expression for $CL(\text{“data”})$ so that the best MDL fitting model can be defined and obtained as its minimizer. It is shown in Appendix B that in the case of competing risks data without masking $CL(\text{“data”})$ can be well approximated by

$$\text{MDL}(A_K) = \sum_{k=1}^{K} \log n_k + \frac{J}{2} \sum_{k=1}^{K} \log(n_k + n_{k+1} \ldots + n_K) - l_{OBS}(\theta), \tag{15}$$

where n_k is the number of observations inside the interval $(a_{k-1}, a_k]$. We propose to select the minimizer of $\text{MDL}(A_K)$ as our MDL–based estimate. Notice that, unlike AICC or BIC, in MDL the penalty for each interval is not the same. Firstly, the penalty for the kth interval is a function of its width n_k. Secondly, from the double summation in the second term of $\text{MDL}(A_K)$, one can see that those "late" intervals (i.e., large k) are penalized more than those "early" intervals (i.e., small k). This agrees with the intuition that stronger penalties (or loosely, more prior information) are required for those "late" intervals because as time passes, more and more items die and hence a smaller amount of information is available for those intervals.

Ideally, one would like to adapt the MDL principle to the situation of masked data. However, due to a technical difficulty given in Appendix B, we decide to use the same criterion for unmasked data. Simulations show that this choice performs better, on average, than AICC and BIC.

3.3 A Fast Splitting Algorithm

Minimizing any one of the above selection criteria with respect to A_K is not a trivial task, as the search space is enormous. Here we describe a simple, fast and yet effective search algorithm for approximating the minimizers of the criteria.

The algorithm starts with fitting a model with only $K = 1$ interval (i.e., no break points) and calculates the corresponding value of the selection criterion being used (i.e., MDL, BIC or AICC). Denote this value as S_1. Then the algorithm adds one break point to the model, or equivalently, split the entire domain into two intervals. The location of this first break point is chosen in the following manner. Among all possible break point locations, if the whole domain is split at this particular break point, it will produce the largest increase (or the smallest decrease) of the likelihood value. To locate such a break point, one could conduct a grid search on $[0, t_{\max}]$, or, in our implementation, we limit the set of all possible break points to be the mid points between any two adjacent observations. To further speed up the algorithm, one could consider say every other mid point. That is, if the set of all mid points are $\{x_1, x_2, \ldots, x_{N-1}\}$, one can consider $\{x_1, x_3, x_5, \ldots, x_{N-1}\}$ instead of all x_i's. Once such a break point is located, the algorithm computes the value of the selection criterion being used. This selection criterion value is denoted as S_2.

The next step of the algorithm is to add one additional break point to the existing

two-interval model; that is, to produce a model with $K = 3$ intervals. This second break point is chosen in a similar manner as before: among all possible splitting locations, it produces the largest increase of the likelihood value after the splitting. After this break point is chosen, the algorithm computes the selection criterion value, S_3. If this computed selection criterion value (S_3) is larger than the value (S_2) obtained with $K = 2$ intervals, then the algorithm stops and the fitted model, among all the fitted models examined so far, that has smallest selection criterion value is taken as the final fitted model. Otherwise, the algorithm continues to add break points to the model, to re-compute and to compare the selection criterion values $S_4, S_5, \ldots$ in a similar fashion as before. The process stops when the selection criterion value S_i increases, and the fitted model with the smallest criterion value S_i is taken as the final fitted model. Some timing figures on the computational speed of this algorithm will be reported in the next section.

It should be noted that for the algorithm defined in Section 2.2 there are certain restrictions on the width of the intervals. More precisely, we need to have for each interval $I_k = (a_{k-1}, a_k]$ and for each cause j at least one item which has failed during I_k and with a failure cause masked in a group that contains j. Such restrictions can be easily incorporated within the fast-splitting algorithm.

4 Examples

4.1 Simulation Study

A simulation study was conducted to empirically evaluate the performances of the above model selection methods. We have used two sets of λ_{jk}'s as our test functions. These

two functions have the same number of failure causes $J = 3$, but have different number of intervals: 3 and 7. The locations of the interval end points and the corresponding values for λ_{jk}'s are listed in Tables 1 and 2.

[Table 1 here]

[Table 2 here]

Altogether three sample sizes $N = (100, 200, 800)$ and three values for the probability p that a masked item is sent to second stage analysis, $p \in \{0.3, 0.6, 1.0\}$ were used (e.g. $p = 1$ means that there are no missing data in the sample). Thus in total the number of different experimental configurations was $2 \times 3 \times 3 = 18$.

For each experimental configuration, 400 simulated data sets were generated, and the following methods were applied to each data set to obtain a fitted model:

- *mdl*: the MDL criterion (15) minimized by the splitting algorithm described in Section 3.3,
- *aicc*: similar to *mdl* but for the AICC criterion (13),
- *bic*: similar to *mdl* but for the BIC criterion (14),
- *f5*: a model with 5 equi-length intervals in $[0, t_{\max}]$. This approach of fixing 5 intervals is a generic approach for situations in which the researcher does not have additional experience with the type of failure data under study and is included in this simulation as a baseline for comparison.

A discrete approximation of the following mean–squared–error (MSE) was used to measure

the quality of the fitted models:

$$\text{MSE} = \sum_{j=1}^{J} \int_0^{t_{\max}} \{\lambda_j(t) - \hat{\lambda}_j(t)\}^2 dt,$$

where the true and known λ_j is $\lambda_j(t) = \sum_{k=1}^{K} \lambda_{jk} 1_k(t)$ and the estimate $\hat{\lambda}_j$ is $\hat{\lambda}_j(t) = \sum_{k=1}^{\hat{K}} \hat{\lambda}_{jk} 1_k(t)$. Boxplots of the log of these MSE values are given in Figures 1 and 2. Paired Wilcoxon tests were also applied to test if the difference between the median MSE values of any two methods is significant or not. The significance level used was 1.25%. If the median MSE value of a particular method is significantly less than the median MSE values of the other three methods, this method will be assigned rank 1. If its median MSE value is significantly less than two but greater than one of the other three, then the method will be assigned a rank 2, and so on for ranks 3 and 4. Methods have insignificantly different MSE values will share the same averaged rank. These paired Wilcoxon rankings are also listed in Figures 1 and 2.

[Figure 1 here]

[Figure 2 here]

One can notice from Figures 1 and 2 that none of the three criteria is uniformly optimal. Table 3 (see below) and also our numerical experience suggest that *aicc* has the tendency to overestimate the number of intervals. When the number of intervals is smaller (i.e. three) *bic* performs the best, with *mdl* lagging not far behind. On the other hand, when the number of intervals is larger (i.e. seven) both *mdl* and *aicc* are doing better than *bic*, with *mdl* being the best.

Surprisingly, the *f5* does not seem to take advantage from additional data. One can see

that as the amount of complete data increases from top to bottom and from left to right *mdl*, *bic* and *aicc* become sensibly more accurate.

[Table 3 here]

The averaged Wilcoxon rankings for *mdl*, *bic*, *aicc* and *f5* are, respectively, 1.72, 1.92, 2.37 and 4. Judging from this measure, it seems that *mdl* should be the preferred method for a researcher without any prior knowledge on the expected number of intervals required by the particular application.

In order to assess the performance of the methods in terms of selecting the correct number of intervals and the correct locations of the end points, we recorded, for those experimental settings associated with $N = 200$ and the test function with three intervals, the number and the end points of the intervals that *mdl*, *bic* and *aicc* selected. Results are summarized in Table 3. The methods *mdl* and *bic* seem to be preferable in this case.

To visually evaluate the quality of the estimates, in Figure 3 we plotted the true λ_{jk}'s for the three-interval test function, together with one representative estimate sampled from the 400 simulations. The number of observations was $N = 400$ and the masking probability was $p = 0.6$. From this figure one can see that both *mdl* and *bic* selected the correct number of intervals, while *aicc* overfitted the data. For *f5*, one can see that it inflated the variance of the estimates confirming the result of Lemma 2.1.

We also applied the above methods to data generated from hazards following a Weibull distribution. In particular we have generated data sets of size $N = 400$ with probability of a second stage analysis $p = 0.6$. In Figure 4 we display the true hazards and the estimate obtained from a typical data set. Not surprisingly, the performances of the three procedures

depend on the gradient of the true hazards. However, it seems fair to say that the *mdl* and *bic* perform well for the three hazards shown. The middle row has a poor fit on the last piece since most of the items die early and few data points are available later in the study. This obviously affects *aicc* and *f5* even more since many of the intervals artificially added in the model will contain very few data points.

[Figure 4 here]

Lastly we report some timing figures. If the true number of intervals is 7 and $N = 200$, our implementation took on average 5 seconds on a Sun Ultra 60 unix workstation for any of the three model selection methods to finish. If $N = 800$, on average it took 17 seconds on the same machine.

4.2 Hard-drive Data

We consider here a real data set analyzed by Flehinger et al. (2002) using a model with Weibull cause-specific hazards. Craiu and Duchesne (2004a) analyze the same data using piecewise constant hazards and their conclusions are very close to those obtained by Flehinger et al. (2002). However, the selection of the intervals was based on a subjective choice, a situation we want to remedy here.

We are interested in the failure causes of a certain sub-assembly of hard-disk drives. Some of these causes are related to particular components (e.g. defective head) but others, such as particle contamination, are not. The analysis does not discriminate among these and simply treats them as causes of failure.

The data consists of $10,000$ hard-drives, of which 172 have failed during an observation

period of 4 years. There are three possible failure causes and for many of the failed items the true cause of death is group masked. There are two masking groups $\{1,3\}$ and $\{1,2,3\}$. The results obtained using *mdl* and *bic* are similar to those of Craiu and Duchesne (2004a) (where the endpoints of the intervals are chosen subjectively) as can be seen from Table 4. The *aic* suggests 8 intervals while the *aicc* suggests 6 intervals. Both models result in inflated variances for the maximum likelihood estimators. The asymptotic standard errors, as measured using the SEM algorithm (Meng and Rubin, 1991), are smaller with the new cut-points $(0, 0.81, 1.58, 3.77, 4)$ compared to the ones obtained previously using the cut-points $(0, 1, 2, 3, 4)$.

5 Conclusions and Further Work

Choosing the end-points of the piecewise cause-specific hazards intervals can play an important part in solving a competing risks problem with or without group masking. We discuss here three possible approaches using the MDL, AICC and BIC criteria. The MDL and BIC seem to be more robust with respect to the number of intervals required for a good approximation. We recommend the use of MDL in situations in which little is known about the number of intervals required.

One can adapt the present work to the case where not all failure causes share the same number of intervals and interval end point locations. It is straightforward to modify the AICC and BIC criteria for this generalization, as the penalty terms in these two criteria are proportional to the number of parameters in the model being fitted. It is also straightforward to derive a corresponding MDL criterion. The modification required is to derive

new expressions for $CL(A_K)$ and $CL(\hat{\theta})$, and the material in Appendix B can be applied to derive such expressions. The fast splitting algorithm discussed in Section 3.3 can also be modified to minimize any of these new criteria. The main idea is, at each time step, instead of adding a new same break point to *all* failure causes, the new algorithm adds one break point to only *one* failure cause. However, this would be a lengthy procedure, as at each time step it requires a lot of comparisons to decide a best break point.

In addition, we would like to expand the present study to fitting splines instead of constant functions on each interval. However, this raises significantly the amount of data required and the complexity of the computation and further research is necessary.

References

Aalen, O. O. (1978), 'Nonparametric inference for a family of counting processes', *Ann. Statist.* **6**, 701–726.

Barrett, A. J. e. a. (1994), 'Marrow transplantation for acute lymphoblastic leukemia at Memorial Sloan-kettering cancer center', *Blood* **74**, 862–871.

Burnham, K. P. & Anderson, D. R. (2002), *Model Selection and Inference: A Practical Information-Theoretic Approach*, second edn, Springer-Verlag New York Inc.

Craiu, R. V. & Duchesne, T. (n.d.*a*), 'Inference based on the EM algorithm for the competing risk model with masked causes of failure', *Biometrika* **91**, 543–558. 2004a.

Craiu, R. V. & Duchesne, T. (n.d.*b*), Using EM and Data Augmentation for the competing risks model, *in* A. Gelman & X. L. Meng, eds, 'Applied Bayesian Modeling and Causal

Inference from an Incomplete-Data Perspective', Wiley. 2004b.

Crowder, M. (2001), *Classical Competing Risks*, Chapman & Hall.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *J. Roy. Statist. Soc. Ser. B* **39**, 1–22.

Dewanji, A. (1992), 'A note on a test for competing risks with missing failure type', *Biometrika* **79**, 855–857.

Dinse, G. E. (1986), 'Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data', *J. Amer. Statist. Assoc.* **81**, 328–335.

Dykstra, R., Kochar, S. & Robertson, T. (1995), 'Likelihood based inference for cause specific hazard rates under order restriction', *J. Multivariate Anal.* **54**, 163–174.

Flehinger, B. J., Reiser, B. & Yashchin, E. (1998), 'Survival with competing risks and masked causes of failures', *Biometrika* **85**, 151–164.

Flehinger, B. J., Reiser, B. & Yashchin, E. (2002), 'Parametric modeling for survival with competing risks and masked failure causes', *Lifetime Data Anal.* **8**, 177–203.

Gaynor, J. J. et al. (1993), 'On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data', *J. Amer. Statist. Assoc.* **88**, 400–409.

Goetghebeur, E. & Ryan, L. (1990), 'A modified log rank test for competing risks with missing failure types', *Biometrika* **77**, 151–164.

Goetghebeur, E. & Ryan, L. (1995), 'Analysis of competing risks survival data when some failure types are missing', *Biometrika* **82**, 821–833.

Hansen, M. H. & Yu, B. (2001), 'Model selection and the principle of minimum description length', *J. Amer. Statist. Assoc.* **96**, 746–774.

Haste, T., Tibshirani, R. & Friedman, J. (2002), *The Elements of Statistical Learning*, Springer, New York.

He, W. & Lawless, J. F. (2003), 'Flexible maximum likelihood methods for bivariate proportional hazards models', *Biometrics* **59**, 837–848.

Hoel, D. G. (1972), 'A representation of mortality data by competing risks', *Biometrics* **28**, 475–488.

Holt, J. D. (1978), 'Competing risk analysis with special reference to matched pair experiments', *Biometrika* **65**, 159–166.

Kalbfleisch, J. D. & Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, 2nd edn, John Wiley & Sons.

Kodell, R. L. & Chen, J. J. (1987), 'Handling cause of death in equivocal cases using the em algorithm (with rejoinder)', *Commun. Statist. A* **16**, 2565–85.

Lagakos, S. W. (1977), 'A covariate model for patially censored data subject to competing causes of failure', *J. Roy. Statist. Soc. Ser. B* **27**, 235–241.

Lapidus, G., Braddock, M., Schwartz, R., Banco, L. & Jacobs, L. (1994), 'Accuracy of fatal motorcycle injury reporting on death certificates', *Accident Anal. and Prevention* pp. 535–542.

Lawless, J. F. (2003), *Statistical Models and Methods for Lifetime Data*, 2nd edn, John Wiley & Sons.

Lee, T. C. M. (2001), 'An introduction to coding theory and the two-part minimum description length principle', *Int. Statist. Rev.* **69**, 169–183.

Lo, S.-H. (1991), 'Estimating a survival function with incomplete cause-of-death data', *J. Multivariate Anal.* **29**, 217–235.

McQuarrie, A. D. R. & Tsai, C.-L. (1998), *Regression and Time Series Model Selection*, World Scientific, Singapore.

Meng, X. L. & Rubin, D. B. (1991), 'Using em to obtain asymptotic variance: The SEM algorithm', *J. Amer. Statist. Assoc.* **86**, 899–909.

Moeschberger, M. L. & David, H. A. (1971), 'Life tests under competing causes of failure and the theory of competing risks', *Biometrics* **27**, 909–933.

Nelson, W. B. (1969), 'Hazard plotting for incomplete failure data', *J. Qual. Technol.* **1**, 27–52.

Prentice, R. L. et al. (1978), 'The analysis of failure times in the presence of competing risks', *Biometrics* **34**, 541–554.

Racine-Poon, A. H. & Hoel, D. G. (1984), 'Nonparametric estimation of the survival function when cause of death is uncertain', *Biometrics* **40**, 1151–8.

Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, World Scientific Publishing Company, Singapore.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Ann. Statist.* **6**, 461–464.

Sun, Y. & Tiwari, R. C. (1997), 'Comparing cumulative incidence functions of a competing-risks model', *IEEE Trans. Reliab.* **46**, 247–253.

Taylor, H. M. (1994), The Poisson-Weibull flaw model for brittle fiber strength, *in* Galambos, J. et al., ed., 'Extreme Value Theory', Amsterdam: Kluwer, pp. 43–59.

A Proof of Lemma 2.1

To show i) write $\hat{\lambda}_{11} = \frac{n_{11}/N}{e_1/N}$ and note that, due to the law of large numbers, the sequences $x_n = n_{11}/N$ and $y_n = e_1/N$ converge almost surely to $\Pr_{M_0}(T \in (0, a_1], C = 1)$ and $E_{M_0}[\min(T, a_1)]$, respectively (the index M_0 signifies that the probability and expectation are computed using the distribution under that model).

From (1) we get

$$\Pr_{M_0}(T \in (0, a_1], C = 1) = \frac{\lambda_1}{\lambda_1 + \lambda_2}(1 - e^{-(\lambda_1+\lambda_2)a_1}) \tag{16}$$

and

$$\begin{aligned} & E_{M_0}[\min(T, a_1)] = E[T\, 1_{\{T \le a_1\}} + a_1 1_{\{T > a_1\}}] \\ = \; & -a_1 e^{(\lambda_1+\lambda_2)a_1} + \int_0^{a_1} e^{-(\lambda_1+\lambda_2)t} dt + a_1 e^{-(\lambda_1+\lambda_2)a_1} = \frac{1 - e^{-(\lambda_1+\lambda_2)a_1}}{\lambda_1 + \lambda_2}. \end{aligned} \tag{17}$$

Under fairly general regularity conditions (16) and (17) imply that $\hat{\lambda}_{11}$ converges almost surely to λ_1. Similar calculations can be done to show that that the same holds for $\hat{\lambda}_1$.

Taking second derivatives of the log-likelihood obtained from (3) one can deduce that the asymptotic variance of $\hat{\lambda}_{11}$, obtained using the observed Fisher information, is n_{11}/e_1^2 while the asymptotic variance of $\hat{\lambda}_1$, is $(n_{11}+n_{12})/(e_1+e_2)^2$. For N large, using i) we have $n_{11}/e_1 \approx \lambda_1 \approx (n_{11}+n_{12})/(e_1+e_2)$ so that the desired result ii) follows.

B Derivation of the MDL Criterion (15)

This appendix outlines the derivation of MDL(A_K). Recall that the goal is to find an expression for

$$CL(\text{“data”}) = CL(A_K) + CL(\hat{\theta}) + CL(\text{“data”}|A_K, \hat{\theta}),$$

and we begin with $CL(A_K)$. Since we limit the break points between different intervals to be a subset of the mid points of any pair of adjacent observations, the width of each interval can be specified with n_k, the number of observations fall within that interval. Thus A_K is completely specified by all n_k's. Using the fact that the code length for an integer I is $\log_2 I$, we have $CL(A_K) = \sum_k CL(n_k) = \sum_k \log_2 n_k$. To calculate $CL(\hat{\theta}) = CL(\hat{\lambda}_{11}) + \ldots + CL(\hat{\lambda}_{JK})$, we apply the following result of Rissanen (1989). If a maximum likelihood estimate is calculated from m data points, then its code length is $\frac{1}{2}\log_2 m$. It can be seen that when there is no masking, for all j, $\hat{\lambda}_{jk}$ is computed from $n_k + \ldots + n_K$ data points; i.e., those items that are still alive. Thus $CL(\hat{\lambda}_{jk}) = \frac{1}{2}\log_2(n_k + \ldots + n_K)$ and hence $CL(\hat{\theta}) = \frac{J}{2}\sum_k \log_2(n_k + \ldots + n_K)$. Lastly, based on Shannon's classical results on information theory, Rissanen (1989) shows that the code length for “data given a fitted

model" amounts to the negative of the conditional log (base 2) likelihood of the data given the fitted model. That is, for our problem, $CL(\text{"data"}|A_K, \hat{\theta}) = -l_{OBS}(\theta)$. Now combining these expressions and changing $\log_2$ to log, we obtain MDL(A_K).

To derive an MDL criterion for masked data, one would need to re-calculate $CL(\hat{\lambda}_{jk})$ for all j and k. This calculation requires the knowledge of the number of items that were used in the computation of $CL(\hat{\lambda}_{jk})$. However, the EM algorithm makes it difficult to track the number of items that were used to estimate each of the λ_{jk}'s. In addition, not all items will have equal weight in (10) since their importance will depend on (9) via (7). We therefore decide to use MDL(A_K) for unmasked data.

Table 1: True λ_{jk} values for the test function with three intervals. The two interval end points are 30 and 50.

j	λ_{j1}	λ_{j2}	λ_{j3}
1	0.0030	0.0200	0.0120
2	0.0045	0.0100	0.0300
3	0.0045	0.0100	0.0300

Table 2: True λ_{jk} values for the test function with seven intervals. The six interval end points are 53, 65, 75, 89, 101 and 173.

j	λ_{j1}	λ_{j2}	λ_{j3}	λ_{j4}	λ_{j5}	λ_{j6}	λ_{j7}
1	0.0013	0.0052	0.0151	0.0001	0.0200	0.0050	0.0500
2	0.0015	0.0081	0.0151	0.0021	0.0300	0.0050	0.0600
3	0.0012	0.0071	0.0161	0.0017	0.0180	0.0060	0.0500

Table 3: Further results for the test function with three intervals and $N = 200$. The left half of the table lists the number of times, out of 400 repetitions, that the number of intervals that a particular method selected. The right half of the table provides the averaged locations of the interval end points for those repetitions that the correct number of intervals were identified. Numbers in the parentheses are estimated standard errors.

		number of intervals				end point 1	end point 2
p	method	2	**3**	4	5+	(at $t = 30$)	(at $t = 50$)
	mdl	54	**285**	54	7	30.06 (0.017)	50.90 (0.296)
0.3	bic	39	**316**	34	11	30.06 (0.016)	50.40 (0.238)
	aicc	0	**24**	66	310	30.00 (0.050)	50.04 (0.432)
	mdl	31	**323**	37	6	30.03 (0.016)	50.19 (0.241)
0.6	bic	20	**346**	28	3	30.04 (0.015)	50.09 (0.219)
	aicc	0	**25**	81	291	30.08 (0.040)	50.24 (0.442)
	mdl	12	**376**	12	0	30.02 (0.015)	50.18 (0.155)
1.0	bic	7	**386**	7	0	30.03 (0.015)	50.14 (0.157)
	aicc	0	**117**	166	117	30.05 (0.027)	50.62 (0.271)

Table 4: Masking probability estimates, Flehinger et al. (2002) hard-drive data. Numbers in parentheses are asymptotic standard errors computed with SEM.

Estimates of the $P_{g\|j}$'s			
Flehinger et al.			
Masking group	$j = 1$	$j = 2$	$j = 3$
$g = \{1, 3\}$	0.412	0	0.446
$g = \{1, 2, 3\}$	0.310	0.469	0.436
Craiu and Duchesne			
Masking group	$j = 1$	$j = 2$	$j = 3$
$g = \{1, 3\}$	0.410 (0.0789)	0	0.445 (0.0563)
$g = \{1, 2, 3\}$	0.308 (0.0766)	0.457 (0.1190)	0.439 (0.0565)
Our estimates			
Masking group	$j = 1$	$j = 2$	$j = 3$
$g = \{1, 3\}$	0.410 (0.0680)	0	0.443 (0.0353)
$g = \{1, 2, 3\}$	0.305 (0.0658)	0.448 (0.0988)	0.442 (0.0356)

Figure 1: Boxplots of the log of the MSE values for the test function with three intervals. The paired Wilcoxon rankings are listed inside the parentheses.

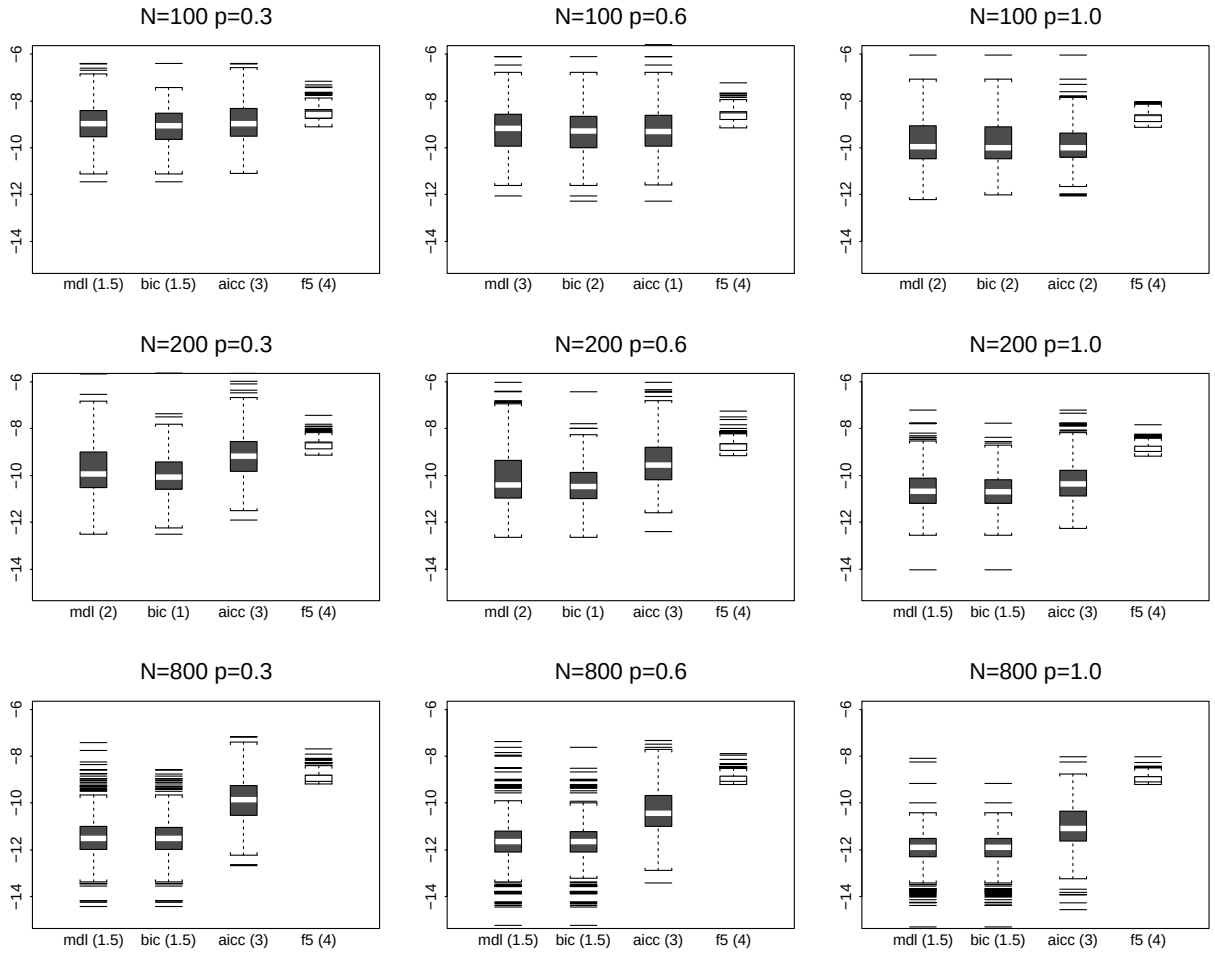

Figure 2: Similar to Figure 1 but for the test function with seven intervals.

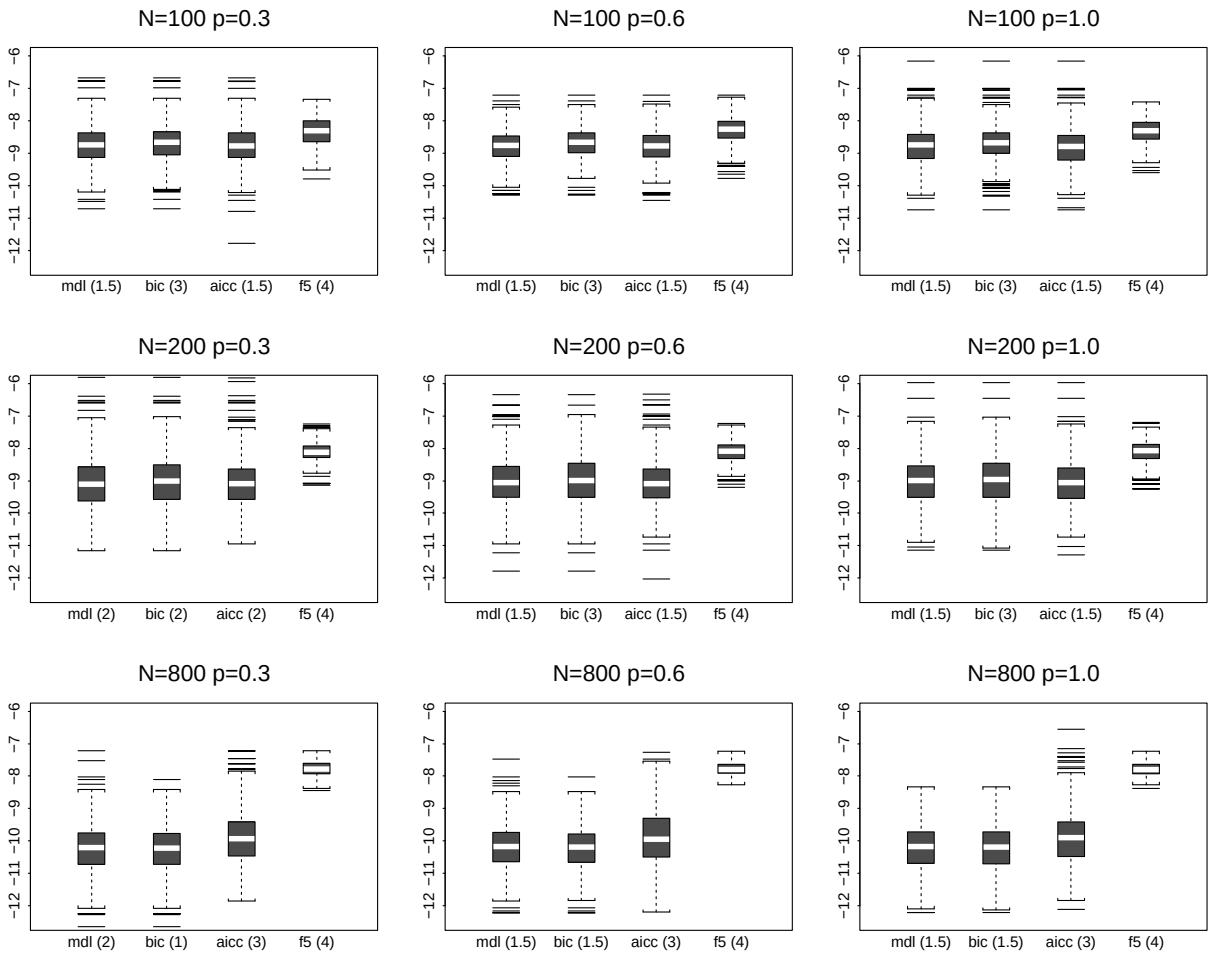

Figure 3: Plots of true (solid lines) and estimated (dotted lines) λ_{jk}'s for the 3-interval test functions.

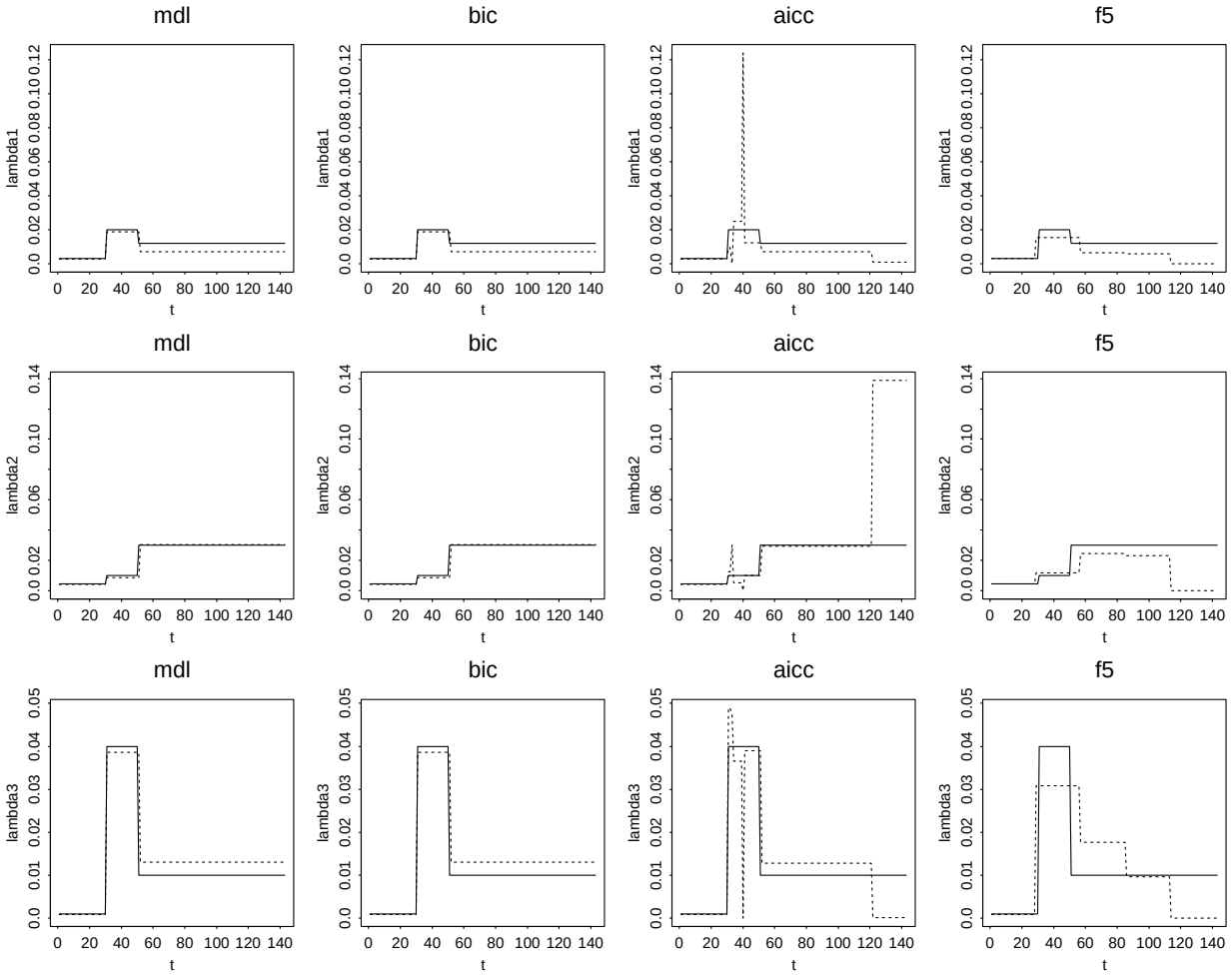

Figure 4: Plots of true (solid lines) and estimated (dotted lines) λ_{jk}'s for Weibull hazards.